Reference: *Biol. Bull.* **194:** 150–160. (April, 1998)

A Major Protein Precursor of Zebra Mussel (*Dreissena polymorpha*) Byssus: Deduced Sequence and Significance

KEVIN E. ANDERSON AND J. HERBERT WAITE*

Chemistry/Biochemistry Department and College of Marine Studies, Newark, Delaware 19716

Abstract. The zebra mussel is a nonindigenous invader of North American lakes and rivers and one of the few freshwater bivalve molluscs having a byssus—a sclerotized organ used by the mussel for opportunistic attachment to hard surfaces. We have sequenced a foot-specific cDNA whose composite protein sequence was deduced from a series of overlapping but occasionally nonidentical cDNA fragments. The overall deduced sequence matches tryptic peptides from a major byssal precursor protein—*Dreissena polymorpha* foot protein 1 (Dpfp1). The calculated mass of Dpfp1 is 49 kDa; but this is known to be extensively hydroxylated and *O*-glycosylated during maturation. Purified native Dpfp1 analyzed using matrix-assisted laser-desorption ionization mass spectrometry with time-of-flight indicates that the protein occurs as at least two size variants with masses of 48.6 and 54.5 kDa. In all probability, the sequence variants reported in this study are related to the larger mass variant. Dpfp1 has a block copolymer-like structure defined by two consensus motifs that are sharply segregated into domains. The N-terminal side of Dpfp1 has 22 tandem repeats of a heptapeptide consensus (P-[V/E]-Y-P-[T/S/δ]-[K/Q]-X); the C-terminal side has 16 repeats of a tridecapeptide motif (K-P-G-P-Y-D-Y-D-G-P-Y-D-K). Both consensus repeats are unique, with some limited homology to other proteins functioning in tension: marine mussel adhesives, plant extensins, titin, and trematode eggshell precursors.

Received 2 December 1997; accepted 10 February 1998.

* To whom correspondence should be addressed. E-mail:hwaite@udel.edu

Abbreviations: DIG, digoxigenin; Dopa, 3, 4-dihydroxyphenylalanine; Dpfp1, *Dreissena polymorpha* foot protein 1; MALDI-TOF, matrix-assisted laser desorption-ionization mass spectrometry with time-of-flight; Mefp1, *Mytilus edulis* foot protein; RACE, rapid amplification of cDNA ends; RT-PCR, reverse transcriptase-polymerase chain reaction; SDS PAGE, sodium dodecyl sulfate polyacrylamide gel electrophoresis.

Introduction

The zebra mussel, *Dreissena polymorpha* (Pallas), is a freshwater bivalve indigenous to the river basins of the Black, Baltic, and Caspian seas. Recently, it was accidentally introduced into one of the Great Lakes, and in less than 10 years, its distribution has expanded into the lakes and rivers of at least a third of the North American continent (Johnson and Padilla, 1996). The economic impact of this expansion has been profound and is due, in large part, to fouling (Roberts, 1990). Zebra mussels foul by attaching opportunistically and in large numbers to a wide variety of surfaces by means of a thread-like structure known as a byssus (Ackerman *et al.*, 1992). In this respect, they resemble marine mussels (Mytilidae), which have adopted a similar strategy.

Zebra mussel byssal threads are fibrous extracellular structures composed largely of proteins, many of which contain the post-translationally modified amino acid 3,4-dihydroxyphenylalanine (Dopa) (Rzepecki and Waite, 1993). Peptidyl Dopa is thus a convenient marker of byssal precursor proteins and is thought to play an important role in adhesion and the maturational cross-linking of byssal threads (Waite, 1990). Three polymorphic Dopa-containing protein families have previously been isolated and partially characterized from zebra mussel foot tissue, the site of byssal protein synthesis and storage. The largest of these proteins, *Dreissena polymorpha* foot protein 1 (Dpfp1), has an apparent molecular weight of 76 kDa and Dopa at levels up to 6.6 mole % (Rzepecki and Waite, 1993). Like many byssal precursors from marine mussels, Dpfp1 features Dopa residues in repeating consensus motifs. Despite this similarity, Dpfp1 is markedly different

from the marine proteins in two respects. First, members of the Dpfp1 family have acidic isoelectric points ranging from 5.3 to 6.5; marine byssal precursors, in contrast, are highly basic—many with pIs exceeding the effective resolving range of available ampholytes. Second, dreissenid byssal precursors, including Dpfp1, are glycosylated with *N*-acetylgalactosamine *O*-linked to serine and threonine residues; there is, however, no evidence for glycosylation in byssal proteins from any marine taxa. It is not known whether these differences reflect two generally valid solutions to the problem of adhesion underwater or represent genuine differences in the requirements for adhesive bond formation in freshwater and marine systems.

Our efforts to determine the complete primary sequence of Dpfp1 by traditional peptide mapping have been thwarted by the repetitive structure and protease-resistance of large regions of the protein (Rzepecki and Waite, 1993). In this study, we report on the complete primary sequence of Dpfp1 deduced using molecular techniques. cDNA sequence data reveal that Dpfp1 is a tandemly repetitive protein composed of two motifs: a novel heptapeptide sequence and a tridecapeptide consensus sequence. Unusually, these motifs are segregated to distinct regions of the protein, a fact which almost certainly has important consequences to the self-assembly of the zebra mussel byssus.

Materials and Methods

RNA extractions

All tissues used in these experiments were excised, immediately frozen in liquid nitrogen, and ground in a mortar chilled to −80°C. Tissue was homogenized in a hand-held glass homogenizer (Kontes, Vineland, NJ), and total RNA was extracted according to the methods of Chomczynski and Sacchi (1987).

Reverse transcriptase (RT)-polymerase chain reaction (PCR) and 5′ rapid amplification of cDNA ends (RACE)

mRNA was purified from total RNA using the Oligotex mRNA spin column kit (Qiagen, Chatsworth, CA). After purification, 1 μg mRNA was reverse transcribed using 20 pmoles of a primer specific to polyA tracts (polyT-LD AGAGAGATTTTTTTTTTTTTTTTVN) with 200 units of MM-LV reverse transcriptase (Superscript II, Gibco-BRL) for 2 h at 37°C in buffer supplied by the manufacturer. The reaction was quenched with 1 ml of 1 × TE, pH 7.5. One percent (v/v) of the resulting first-strand cDNA was amplified with the polymerase chain reaction (PCR) using degenerate oligonucleotide primers based on the previously determined (Rzepecki and Waite, 1993) amino acid sequence of the N-terminus of Dpfp1 (Dp1.N(+) GGIACITAYGAYTGGACNGA) and an internal peptide (Dp1.A(−) TTRTCRTAIGGICCRTCRTA). Each 50-μl reaction contained 0.25 m*M* of each dNTP, 100 pmoles of each primer, and 2.5 units of *Taq*2000 polymerase (Stratagene, La Jolla, CA), in a buffer containing 10 m*M* Tris-Cl, 1.5 m*M* $MgCl_2$, 75 m*M* KCl, and 15 m*M* $(NH_4)_2SO_4$. Samples were initially denatured at 95°C for 4 min 30 s followed by 30 cycles of amplification as follows: 95°C for 30 s, 50°C for 30 s, and 72°C for 2 min. A final extension for 5 min at 72°C was carried out to ensure addition of 3′ A overhangs. The resulting amplification product was ligated into the pCRII vector (Invitrogen, San Diego, CA) according to manufacturer's instructions. The insert from the newly constructed plasmid, pDP1.NA, was sequenced on both strands using vector-specific and degenerate oligonucleotide primers.

5′ RACE was performed to obtain cDNA sequence data upstream of the region coding for the N-terminus (Frohman *et al.*, 1988) and to independently establish the cDNA sequence of the N-terminus. All reactions were performed using reagents contained in the 5′ RACE System V2.0 (Life Technologies, Bethesda, MD) according to manufacturer's instructions. Briefly, 1 μg of *D. polymorpha* foot tissue total RNA was reverse-transcribed using a gene-specific primer (Dp1.GSP1(−) TATTTTGTAGGAGTGGG). The purified first-strand cDNA was tailed with dCTP, and PCR was performed using the supplied abridged anchor primer (GGCCACGCGTCGACTAGTACGGGIIGGGIIGGGIIG) and Dp1.GSP1(−). Each 50-μl reaction contained 0.25 m*M* of each dNTP and 20 pmoles of each primer in 1× PCR buffer (Life Technologies, Bethesda, MD) supplemented with 2 m*M* $MgCl_2$. Samples were denatured at 95°C for 4 min 30 s and equilibrated to 72°C. Two-and-one-half units of *Taq*2000 polymerase were added and amplification for 25 cycles was performed under the following conditions: 95°C for 30 s, 42°C for 30 s, and 72°C for 30 s. A final 5-min extension was performed at 72°C. A second round of PCR was performed using AAP and a nested gene-specific primer (Dp1.GSP2(−) TTGTTGTATAGTTCGGAATTTTAG). The reaction volume and component concentrations were as outlined in the previous reaction. Samples were initially denatured at 95°C for 4 min 30 s followed by 30 cycles of amplification as follows: 95°C for 30 s, 42°C for 30 s, and 72°C for 60 s. A final extension for 5 min at 72°C was carried out to ensure addition of 3′ A overhangs. The resulting amplification products were cloned into the pGEM-T vector (Promega, Madison, WI) according to manufacturer's instructions. The insert from the newly constructed plasmid, pDP1.5′UTA, was sequenced on both strands using gene-specific primers.

Probe synthesis and cDNA library screening

Two probes were created in this experiment to screen a *D. polymorpha* foot tissue cDNA library (Eddington,

1996). A digoxigenin (DIG)-labeled antisense RNA probe (probe #1) was generated from DdeI-digested pDP1.NA using T7 polymerase and the DIG-RNA labeling kit (Boehringer-Mannheim) according to manufacturer's instructions. A DIG-labeled double-stranded DNA probe (probe #2) spanning the 5′ untranslated region of Dpfp1 and the first 172 nt coding for the mature protein was generated using the PCR DIG probe synthesis kit (Boehringer-Mannheim) according to manufacturer's instructions. pDP1.5′UTA was used as a template for this reaction, and a primer specific to the 5′ untranslated region of Dpfp1 (Dp1.5′ UT(+) ATACTTCAGAGCATCAACCAA) and Dp1.GSP1(−) were used as primers. Both probes were individually incorporated at a concentration of 100 ng/ml into standard hybridization buffer + 50% formamide (5× SSC, 1% Blocking buffer (Boehringer-Mannheim), 0.1% (w/v) sarcosyl, 0.02% (w/v) SDS, 50% formamide (v/v)). Hybridizations were carried out at 60°C (probe #1) or 42°C (probe #2). Stringency washes for both probes were conducted with 0.1× SSC/0.2% (w/v) SDS at 68°C.

One million plaques generated from a λZAP-Express cDNA library (Stratagene, La Jolla, CA) were doubly screened with probes #1 and #2. No plaques positive for probe #2 were detected, suggesting that a full-length clone of Dpfp1 was not present in this library. Forty plaques positive for probe #1 were cored, eluted in SM buffer (100 m*M* NaCl, 50 m*M* Tris-Cl pH 7.5, 8 m*M* $MgSO_4$, 0.1% gelatin), and tested for insert size by PCR using vector-specific primers flanking the cDNA insert. After secondary screening, cDNA from the plaque bearing the largest insert was rescued as a phagemid using the ExAssist interference-resistant helper phage kit (Stratagene, La Jolla, CA) and sequenced using the nested deletion technique (see below).

Nested deletions

Nested deletions were performed using the double-stranded nested deletion kit (Pharmacia Biotech, Piscataway, NJ). In each case, 5 μg of template was doubly digested with EcoRI and PstI, and the restriction enzymes were heat inactivated. Digested clones were precipitated in ethanol and resuspended in a buffer containing 1.5 *M* potassium acetate, 37.5 m*M* Tris-acetate pH 7.6, 15 m*M* magnesium acetate, 750 μ*M* β-mercaptoethanol, and 15 μg/ml bovine serum albumin (BSA). A 2-μg sample of each digest was used for digestion with Exonuclease III. The reactions were carried out at 23°C and aliquots taken every 5 min. All clones yielding deletions larger than the size of the empty vector were ligated, transformed into XL1-Blue MRF′ cells (Stratagene, La Jolla, CA), purified, and sequenced using a vector-specific primer.

RNA dot blots

Ten micrograms of total RNA separately extracted from *D. polymorpha* foot, adductor mussel, mantle, and gill tissue were diluted in an equal volume of RNA dilution buffer (water: 20× SSC: formaldehyde; 5:3:2) and spotted onto a positively charged nylon membrane (MSI, Westboro, MA). The membrane was hybridized to either probe #1 as described above or to an actin-specific double-stranded DIG-labeled DNA probe (Patwary *et al.*, 1996). Hybridization with actin specific probe was performed at 37°C with a stringency wash using 0.5× SSC/0.1% (w/v) SDS at 68°C.

Northern hybridizations

Three micrograms of foot tissue mRNA were subjected to formaldehyde/agarose gel electrophoresis according to Sambrook *et al.* (1989). RNA was transferred onto a positively charged nylon membrane and hybridized overnight with probe #1.

Mass analysis of native Dpfp1

Native Dpfp1 was purified from the foot of adult zebra mussels according to Rzepecki and Waite (1993). The mass of the native protein was determined by matrix-assisted laser desorption-ionization mass spectrometry with time-of-flight (MALDI-TOF) using a PerSeptive Biosystems Voyager model in the positive ion mode and delayed extraction. A 20-μ*M* solution of Dpfp1 in 0.1% acetic acid was mixed with three volumes of a saturated sinapinic acid solution (40% acetonitrile/0.1% TFA); 2 μl of the resulting mixture (10 pmoles Dpfp1) was placed on a sample plate and allowed to air dry. The sample was inserted into a vacuum chamber (1×10^{-7} torr) and the spectra generated from 256 pulses of a 337-nm laser were averaged. The acceleration voltage was 25,000 with a 90% grid voltage and a guidewire setting of 0.1%.

Results

RNA dot blots and Northern hybridizations

The tissue specificity of Dpfp1 is demonstrated in Figure 1. RNA dot blots show that Dpfp1 mRNA transcripts were detected only in total RNA extracts from foot tissue and not in extracts from gill, adductor muscle, or mantle tissue. Identical dot blots hybridized to an actin-specific probe were positive for all tissue types although the strength of the signal varied considerably between tissue types (data not shown). These results are consistent with data obtained from other marine byssal precursor proteins (Inoue *et al.*, 1995, 1996a; Coyne *et al.*, 1997; Qin *et al.*, 1997) and support the hypothesis that Dpfp1 plays a role

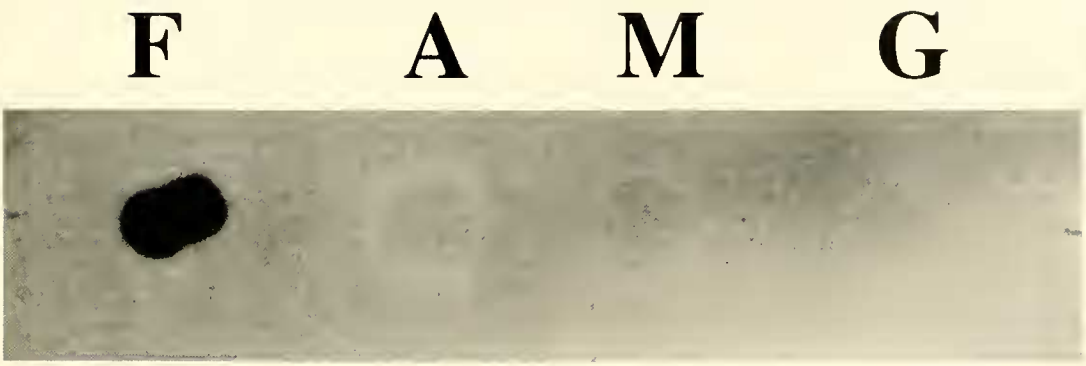


Figure 1. Total RNA dot blots of *Dreissena polymorpha* foot (F), adductor muscle (A), mantle (M), and gill (G) tissue hybridized to a digoxigenin-labeled RNA probe specific to Dpfp1. The probe was hybridized to 1 μg of total RNA from each tissue.

as a byssal structural protein. Northern blots of foot tissue mRNA indicated that Dpfp1 transcripts range in size from 1200 b to 1500 b, suggesting the presence of size variants (Fig. 2).

Dpfp1 cDNA sequence

In Figure 3 the aligned nucleotide sequence data obtained from 5′ RACE, RT-PCR with degenerate oligonucleotide primers, and from the largest cDNA clone isolated are presented. Each sequence differs slightly from the other, and therefore the consensus sequence generated from this alignment does not represent any single Dpfp1 sequence. It is likely that differences in the data sets reflect the existence of Dpfp1 variants rather than errors introduced during amplification, because each set of PCR sequence data was determined from at least two independently amplified samples. The combined transcript is 1481 bp in length and contains an open reading frame of 1332 bp coding for a protein of 443 amino acids. Included in the transcript is a start codon at nucleotide position 36 and two overlapping canonical polyadenylation signals (Kozak, 1986) at nucleotide positions 1464 and 1468. The calculated molecular weight of the deduced primary sequence is 49 kDa, with a predicted isoelectric point of 5.29.

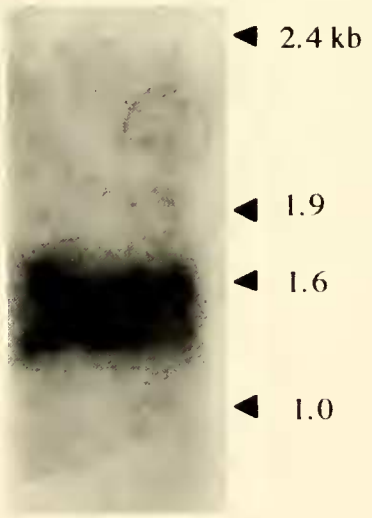
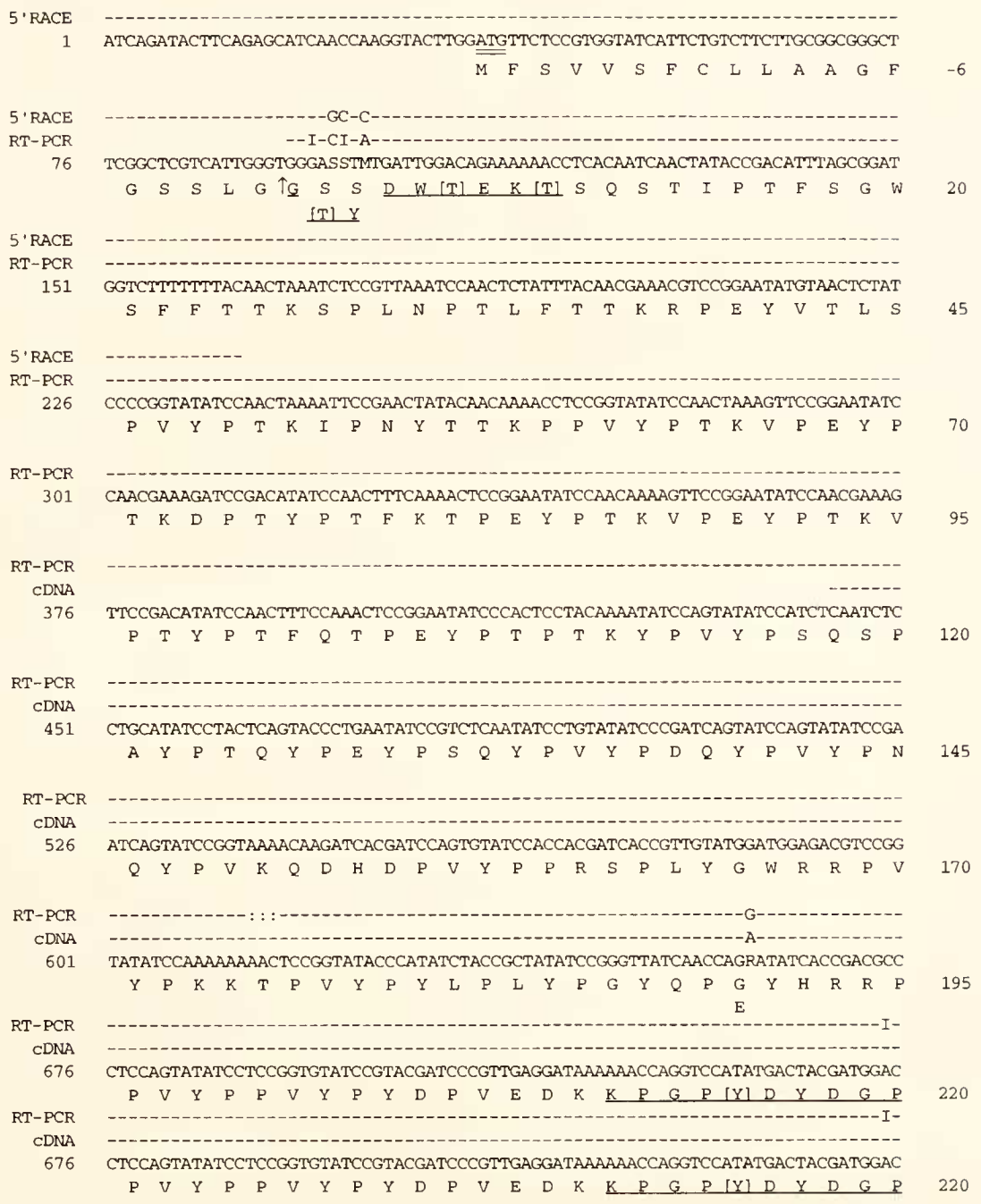


Figure 2. Northern blot of *Dreissena polymorpha* foot tissue mRNA hybridized to a digoxigenin-labeled RNA probe specific to Dpfp1; 3 μg of foot tissue mRNA was used.

The first 19 amino acids code for a putative signal peptide that conforms to the rule of von Heijne (1985). Computer-based modeling of signal peptide cleavage (Nielsen *et al.*, 1997) correctly predicts cleavage of the signal peptide preceding the previously determined N-terminal glycine residue of the mature protein (Rzepecki and Waite, 1993). The N-terminus of Dpfp1, as coded for by sequences generated using 5′ RACE, differs from the previously reported N-terminal sequence (Rzepecki and Waite, 1993) in that it substitutes serine residues for threonine at position #2, tyrosine at position #3, and aspartic acid at position #10. None of the three independently generated 5′ RACE clones exactly coded for the previously reported N-terminus of Dpfp1. N-terminal sequence data generated with degenerate oligonucleotide primers more closely resemble the previously reported N-terminal sequence but also substitute serine for aspartic acid at position #10. It is not possible to determine from these data if the N-terminal sequence deduced from cDNAs generated via degenerate oligonucleotide primers reflects a genuinely different N-terminus or is simply an artifact forced by the primers used during amplification.

The N-terminal 38 amino acids of the mature protein are relatively enriched in threonine and serine residues and quickly give way to a tandemly repeating heptapeptide. This generally basic motif (P-[V/E]-Y-P-[T/S]-[K/Q]-X) is repeated 22 times in the N-terminal half of Dpfp1


5'RACE
1
ATCAGATACTTCAGAGCATCAACCAAGGTACTTGGATGTTCTCCGTGGTATCATTCTGTCTTCTTGCGGCGGGCT
M F S V V S F C L L A A G F -6
5'RACE
RT-PCR
76
TCGGCTCGTCATTGGGTGGGASSTMTGATTGGACAGAAAAAACCTCACAATCAACTATACCGACATTTAGCGGAT
G S S L G G S S D W [T] E K [T] S Q S T I P T F S G W 20
[T] Y
5'RACE
RT-PCR
151
GGTCTTTTTTTACAACTAAATCTCCGTTAAATCCAACTCTATTTACAACGAAACGTCCGGAATATGTAACTCTAT
S F F T T K S P L N P T L F T T K R P E Y V T L S 45
5'RACE
RT-PCR
226
CCCCGGTATATCCAACTAAAATTCCGAACTATACAACAAAACCTCCGGTATATCCAACTAAAGTTCCGGAATATC
P V Y P T K I P N Y T T K P P V Y P T K V P E Y P 70
RT-PCR
301
CAACGAAAGATCCGACATATCCAACTTTCAAAACTCCGGAATATCCAACAAAAGTTCCGGAATATCCAACGAAAG
T K D P T Y P T F K T P E Y P T K V P E Y P T K V 95
RT-PCR
cDNA
376
TTCCGACATATCCAACTTTCCAAACTCCGGAATATCCCACTCCTACAAAATATCCAGTATATCCATCTCAATCTC
P T Y P T F Q T P E Y P T P T K Y P V Y P S Q S P 120
RT-PCR
cDNA
451
CTGCATATCCTACTCAGTACCCTGAATATCCGTCTCAATATCCTGTATATCCCGATCAGTATCCAGTATATCCGA
A Y P T Q Y P E Y P S Q Y P V Y P D Q Y P V Y P N 145
RT-PCR
cDNA
526
ATCAGTATCCGGTAAAACAAGATCACGATCCAGTGTATCCACCACGATCACCGTTGTATGGATGGAGACGTCCGG
Q Y P V K Q D H D P V Y P P R S P L Y G W R R P V 170
RT-PCR
cDNA
601
TATATCCAAAAAAAACTCCGGTATACCCATATCTACCGCTATATCCGGGTTATCAACCAGRATATCACCGACGCC
Y P K K T P V Y P Y L P L Y P G Y Q P G Y H R R P 195
E
RT-PCR
cDNA
676
CTCCAGTATATCCTCCGGTGTATCCGTACGATCCCGTTGAGGATAAAAAACCAGGTCCATATGACTACGATGGAC
P V Y P P V Y P Y D P V E D K K P G P [Y] D Y D G P 220
RT-PCR
cDNA
676
CTCCAGTATATCCTCCGGTGTATCCGTACGATCCCGTTGAGGATAAAAAACCAGGTCCATATGACTACGATGGAC
P V Y P P V Y P Y D P V E D K K P G P [Y] D Y D G P 220

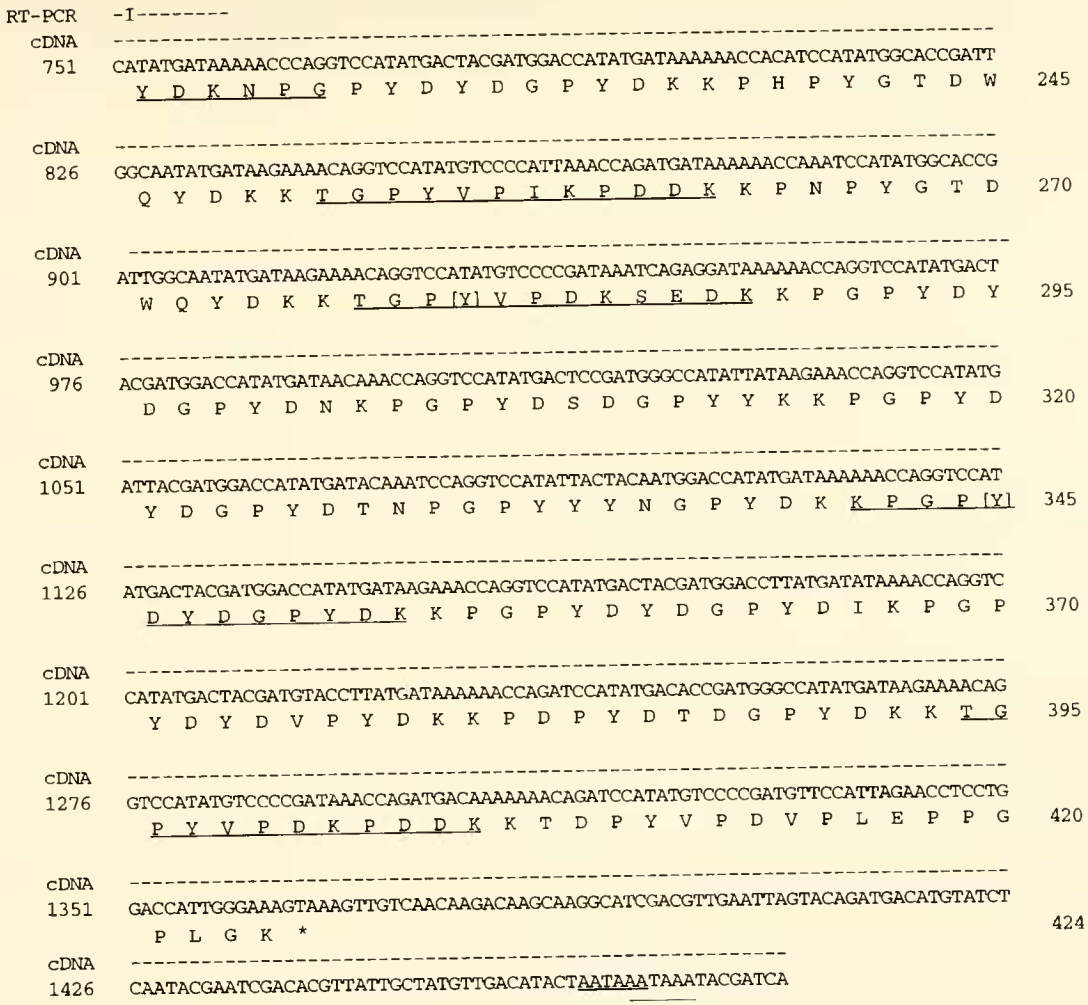


Figure 3. Composite sequence of Dpfp1 obtained by the alignment of sequences derived from 5′ RACE, RT-PCR with degenerate oligonucleotide primers, and the largest partial cDNA clone isolated. Colons (:) are gaps inserted for alignment purposes; dashes (-) indicate agreement with the consensus sequence. Positions coded for by inosine are marked as I. The consensus sequence in these regions reflects the sequence of the non-inosine-containing strand. Underlined nucleotide sequences contain the canonical polyadenylation signal (Kozak, 1986), with doubly underlined nucleotides indicating the position of the start codon. An asterisk (*) indicates a stop codon. The deduced amino acid sequence is presented with residues represented by the single letter code. ↑ indicates predicted cleavage position of signal peptide (von Heijne, 1985; Nielsen *et al.*, 1997). Underlined amino acid residues have been mapped to tryptic peptides from Rzepecki and Waite (1993). Bracketed amino acid residues represent post-translational modifications mapped from Rzepecki and Waite (1993). [T] indicates a glycosylated threonine residue; [Y] represents a Dopa residue in the mature protein.

with some variation, particularly at position #7 of the consensus sequence; however, proline residues at positions #1 and #4 and tyrosine residues at position #3 are highly conserved (Fig. 4). RT-PCR data differ from cDNA clone data in this region of the transcript by omission of threonine 175 and by a G190E substitution resulting from a transversion at nucleotide position 661.

The C-terminal half of Dpfp1 is dominated by the previously reported 13 amino acid consensus sequence: K-P-G-P-Y-D-Y-D-G-P-Y-D-K (Rzepecki and Waite, 1993). This acidic sequence is found tandemly repeated 16 times with only slight variations from the consensus (Fig. 4). The deduced amino acid composition of the composite Dpfp1 sequence, without signal peptide sequence agrees well with that of native Dpfp1 (Table I), suggesting that the composite sequence described above is representative of Dpfp1 mRNAs present in zebra mussel foot tissue. Examination of codon usage for Dpfp1 (Table II) reveals a significant degree of codon bias in amino acids that occur in conserved positions of the above-mentioned consensus sequences (*e.g.*, P, Y, D, K, T, G).

Mass analysis of native Dpfp1

MALDI-TOF analysis of native Dpfp1 indicates that the purified protein is represented by two major mass variants. The lighter of the two variants has a mass $[M + H^+]^+ = 48.6$ kDa, whereas in the heavier variant, $[M + H^+]^+ = 54.5$ kDa. No peaks were detected in the 60–80 kDa range.

Discussion

The primary structure of Dpfp1, deduced from overlapping cDNAs, represents the first complete sequence for a dreissenid byssal protein and an important advance in understanding the attachment strategy of the zebra mussel. Two observations suggest that the composite sequence generated from these data sets is likely to resemble full-length transcripts for Dpfp1. First, the size of the composite sequence (1481 bases) closely matches the size of the largest Dpfp1 transcript as determined by Northern blots of zebra mussel foot tissue mRNA hybridized to a Dpfp1-specific probe. Second, the deduced amino acid

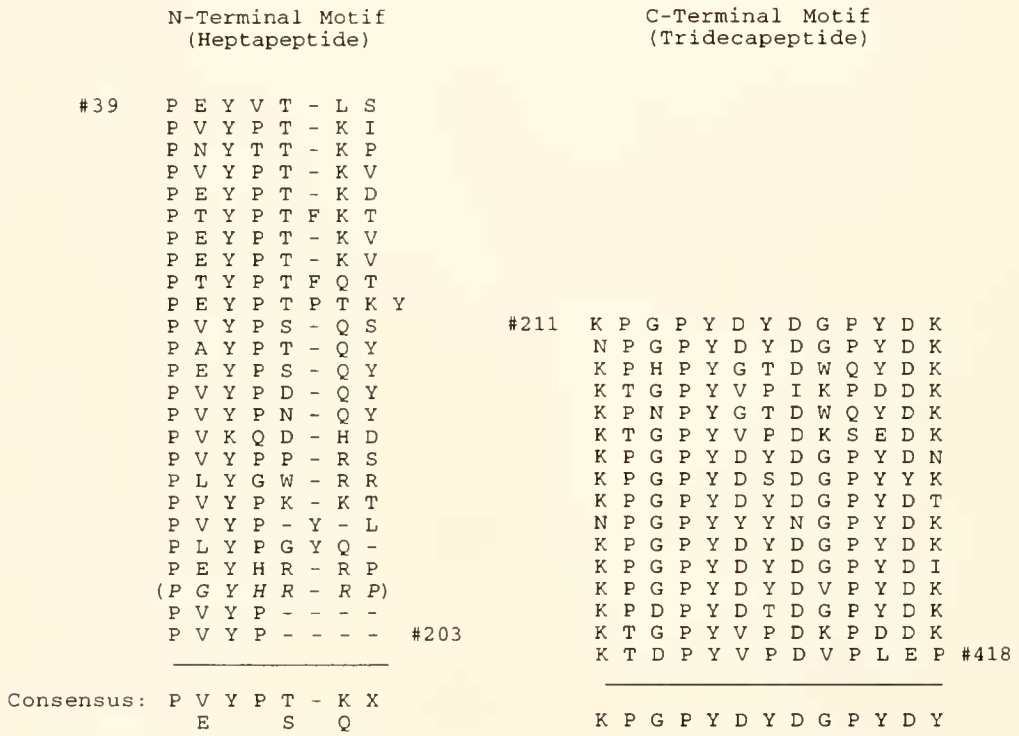


Figure 4. Aligned motifs of Dpfp1. The consensus sequences of Dpfp1 are presented as derived from the tandemly repeating motifs of the deduced primary sequence. The sequences comprising each motif are presented contiguously, and amino acids in the consensus sequence occur in the majority of aligned motifs at their respective positions. Numbers at the beginning and end of each motif represent the position of this sequence within the deduced amino acid sequence for Dpfp1 as presented in Figure 3. The motif in parentheses occurs in RT-PCR derived sequences in place of the preceding cDNA motif. Amino acids are represented by their single letter codes, and a dash (-) indicates a gap inserted for alignment purposes.

Table I

Amino acid composition of deduced and native Dpfp1

Amino acid	Native	Deduced
Asx	136.7	134.8
Thr	75.0	82.7
Ser	34.4	33.1
Glx	70.1	52.0
Pro	238.6	234.0
Gly	76.5	68.6
Ala	7.9	2.4
Val	50.4	52.0
Met	0.7	0.0
Ile	9.9	9.5
Leu	20.4	18.9
Dopa	66.6	N.D.
Tyr	84.5	165.5
Phe	9.6	14.2
His	5.1	7.1
Lys	94.8	99.3
Arg	17.0	14.2
Trp	1.8	11.8
Total:	1000.0	1000.0

The amino acid composition of deduced Dpfp1 is determined excluding signal peptide residues, and that of native Dpfp1 is from Rzepecki and Waite (1993). All values are in residues per thousand residues.

composition of the composite sequence, excluding the signal peptide, closely matches the composition of native Dpfp1 as reported in Rzepecki and Waite (1993).

Purified native Dpfp1 was subjected to MALDI-TOF analysis to resolve the conflict between the apparent and cDNA-deduced mass estimates. SDS-PAGE of native Dpfp1 established that the purified protein migrates as a doublet with apparent molecular masses of 65 and 76 kDa (Rzepecki and Waite, 1993). However, the deduced mass of Dpfp1 of 49 kDa (this work), even allowing for an additional 6.5 kDa contributed by post-translational glycosylation and hydroxylation (Rzepecki and Waite, 1993), is difficult to reconcile with the empirically determined apparent masses. According to MALDI-TOF mass spectrometric analysis, Dpfp1 exists primarily as a doublet (48.6 and 54.5 kDa) with no visible components above 60 kDa. The mass of the larger variant is in excellent agreement with the deduced mass of Dpfp1 after addition of post-translational modifications. The smaller variant may represent unmodified Dpfp1 or possibly a fully modified variant coded for by one of the smaller Dpfp1 transcripts detected during Northern blot analysis of mRNA from zebra mussel foot tissue (Fig. 2). This observation confirms that Dpfp1, like many other byssal precursor proteins (see Coyne *et al.*, 1997; Qin *et al.*, 1997; Taylor *et al.*, 1996; Papov *et al.*, 1995), migrates anomalously during SDS-PAGE.

In previous studies, isoelectric focusing of purified Dpfp1 suggested the presence of at least 10 electrophoretic variants in the polymorphic family (Rzepecki and Waite, 1993). These multiple bands may reflect differences in the primary structure of Dpfp1 variants, nonuniform post-translational modification of one or more forms of the protein, or both. At least some of the variation must arise from differences in primary structure since the N-terminus of Dpfp1 exhibited heterogeneity at two

Table II

Codon usage in Dpfp1

Amino acid	Codon	#	Amino acid	Codon	#	Amino acid	Codon	#	Amino acid	Codon	#
F	TTT	4	L	CTT	2	I	ATT	2	V	GTT	5
F	TTC	5	L	CTC	0	I	ATC	0	V	GTC	4
L	TTA	2	L	CTA	4	I	ATA	2	V	GTA	12
L	TTG	3	L	CTG	0	M	ATG	1	V	GTG	2
S	TCT	6	P	CCT	12	T	ACT	13	A	GCT	0
S	TCC	3	P	CCC	7	T	ACC	4	A	GCC	0
S	TCA	6	P	CCA	58	T	ACA	15	A	GCA	1
S	TCG	1	P	CCG	22	T	ACG	3	A	GCG	2
Y	TAT	57	H	CAT	1	N	AAT	5	D	GAT	39
Y	TAC	12	H	CAC	2	N	AAC	3	D	GAC	9
*	TAA	1	Q	CAA	8	K	AAA	39	E	GAA	9
*	TAG	0	Q	CAG	3	K	AAG	6	E	GAG	2
C	TGT	1	R	CGT	2	S	AGT	0	G	GGT	14
C	TGC	0	R	CGC	1	S	AGC	2	G	GGC	4
*	TGA	0	R	CGA	2	R	AGA	1	G	GGA	11
W	TGG	5	R	CGG	0	R	AGG	0	G	GGG	3

Amino acids are represented by their single letter codes, and an asterisk (*) indicates a stop codon. These data are compiled from the composite sequence of Dpfp1 presented in Figure 3. Where discrepancies exist in the consensus sequence, the cDNA and 5′ RACE data are used to determine codon usage at that position.

positions (#2 and #8, Fig. 3) (Rzepecki and Waite, 1993). The nucleotide sequences presented in Figure 3 suggest the existence of at least two of these variants. Differences between these variants in regions of cDNA overlap are limited to the deletion of a single codon in the RT-PCR data and a single transversion resulting in an amino acid substitution in one of the heptapeptide sequences.

An examination of the codon usage data (Table II) indicates that compositionally dominant amino acids are predominantly coded for by half of the potentially available codons for these residues. This is especially true of proline, tyrosine, aspartic acid, lysine, threonine, and glycine residues, which together account for almost 75% of the amino acid composition of Dpfp1. The pattern of codon bias in compositionally dominant residues has also been noted in other marine byssal precursor proteins—notably Mefp1 (Filpula *et al.*, 1990), Mgfp1 (Inoue and Odo, 1994), Mcfp1 (Inoue *et al.*, 1996b), and, to a lesser extent, Mgfp2 (Inoue *et al.*, 1995)—and may reflect a need to express byssal structural proteins rapidly in response to developmental cues and changing environmental conditions. It is well established that in bacterial systems, codon bias is positively correlated with the rates of gene expression (Robinson *et al.*, 1984; Varenne *et al.*, 1984; Sorensen *et al.*, 1989), presumably through selection of codons that recognize the most abundant isoaccepting tRNAs for a given amino acid. Precedence for this hypothesis can also be found among highly expressed genes in multicellular organisms such as *Drosophila melanogaster*, whose chorion genes, important eggshell components known to be highly expressed during egg development (Kafatos *et al.*, 1987), also exhibit significant codon bias (Akashi, 1994). Such a hypothesis has also been advanced to explain observed codon bias in the highly expressed silk fibroin heavy chain of the silk moth, *Bombyx mori* (Mita *et al.*, 1994).

More than 80% of the deduced primary amino acid sequence of Dpfp1 is composed of tandemly repeated and segregated motifs: one is a heptapeptide; the other, a tridecapeptide consensus motif that coincides with peptides sequenced previously (Rzepecki and Waite, 1993). The occurrence of two relatively short tandemly repeating motifs in Dpfp1 is consistent with its proposed role as a byssal structural protein. However, the absence of data on the distribution of Dpfp1 within the byssus makes it difficult to assign a specific role at this time. The repetitive nature of Dpfp1 is shared by many of the structural proteins of marine byssi. Two of three characterized Dopa-containing byssal proteins in *Mytilus* are known to be composed almost entirely of tandem repeats. Mefp1, a 110-kDa protein thought to play a role as a cuticular lacquer in the byssus of *M. edulis*, is dominated by non-segregated hexa- and decapeptide repeats (Filpula *et al.*, 1990; Waite *et al.*, 1985; Laursen, 1992). Mgfp2, a 49-kDa plaque-specific protein of *M. galloprovincialis*, is largely composed of larger, epidermal growth factor-like repeats (Inoue *et al.*, 1995).

The N-terminal half of Dpfp1 is dominated by a heptapeptide motif that is repeated 22 times with some variation, particularly at position #7 of the consensus sequence. Variability notwithstanding, the spacing of proline and tyrosine residues is well conserved, suggesting that these amino acids play an important functional role in the motif. No tryptic peptides exactly matching the deduced primary sequence could be mapped to this part of the protein; however, a fragment of one tryptic peptide (tryptic peptide #13 in fig. 6 of Rzepecki and Waite, 1993) containing the subsequence S-P-L-Y-G-W . . . is found to bridge two of the heptapeptide repeats. Although the tyrosine in this sequence is efficiently converted to Dopa, the amino acid composition of residual undigested Dpfp1 suggests that, as a whole, this region contains relatively little Dopa (Rzepecki, pers. comm.).

Given the frequency of lysine and arginine in the heptapeptide repeat region, the resistance of the repeat to cleavage by trypsin is intriguing. An examination of the deduced primary sequence indicates that K-P or R-P sequences cannot be the basis for this resistance. Interestingly, lysine and arginine residues in this domain frequently occur adjacent to threonine and serine residues. That observation, coupled with the detection of high levels of threonine and *N*-acetylgalactosamine in partially digested tryptic peptides (Rzepecki and Waite, 1993), leads to the hypothesis that Arg and Lys are protected from trypsin cleavage by adjacent glycosylated amino acids. A similar protection appears to be imparted by glycosylated residues in an extensin-like glycoprotein from *Volvox carteri* (Ertl *et al.*, 1992).

The N-terminal half of Dpfp1 differs significantly from the C-terminal domain with its repeated 13 amino acid motif (Fig. 4). Previous peptide data (Rzepecki and Waite, 1993) and the deduced sequence of Dpfp1 are consistent with the hypothesis that glycosylation is more extensive in the N-terminal region of the protein, whereas hydroxylation of tyrosine to Dopa occurs more frequently in the remaining C-terminal portion. Additionally, the average isoelectric point of Dpfp1 in the region occupied by the heptapeptide is moderately basic (pI = 8.7), whereas the C-terminal domain is quite acidic (pI = 4.7). These divergent characteristics suggest that the segregation of motifs plays a significant role in the architectural design of the zebra mussel byssus. Recently, two byssal structural proteins from *M. edulis* have also been shown to be composed of "block copolymer"-like domains. Both proteins have a central collagenous core flanked by sequences resembling either elastin (Coyne *et al.*, 1997) or silk fibroin (Qin *et al.*, 1997). The distribution of these proteins can

A. Heptapeptide comparisons

Protein	*Consensus Sequence*	*Repeats*	*Ref.*
Dpfp1	P V Y P - T - K - X	22	a
Mefp1	P S Y P P T Y K A K	75	b
Soybean PRP	P V Y - - - - K - P	43	c
Titin PEVK	P V - P - X - K	27	d

a) present data; b) Laursen (1992); c) Hong et al (1987) ; d) Labeit and Kolmerer (1995)

B. Tridecapeptide comparison

Protein	*Consensus Sequence*	*Repeats*	*Ref.*
Dpfp1	K P G P Y D Y D G P Y D K	15	a
Eggshell TRP	Y - G - Y D K Y G - Y D K	27	e

a) present data; e) Wells and Cordingly (1992)

Figure 5. Comparison of the heptapeptide (A) and tridecapeptide (B) motifs of Dpfp1 with a variety of other repetitive proteins. Shaded residues highlight identities; gaps denoted by dashes (-) are included to maximize alignment. X denotes any amino acid residue.

be used to account for the heterogeneous mechanical properties of byssus in *M. edulis* (Qin and Waite, 1995).

Although the consensus motifs of Dpfp1 do not have strong homologies with any known structural proteins, they do share some features with other proteins containing tandem repeats—*i.e.*, marine adhesives (Laursen, 1992), extensin-like proteins from plants (Kieliszewski and Lamport, 1994), and a trematode eggshell protein (Wells and Cordingley, 1992) (Fig. 5). The β-turn (Pro Val) and lysine of the heptapeptide are prominent in extension (soybean PRP) (Hong *et al.*, 1987) and adhesive protein (Waite *et al.*, 1985). In addition, although not a repeating sequence, the PEVK domain of titin, a protein of skeletal muscle, contains at least 27 occurrences of the motif $PVPX_nK$ in which X_n can be from one to three amino acids long (Labeit and Kolmerer, 1995). The tridecapeptide of Dpfp1, in contrast, shares the repeated proximity of YD with a trematode eggshell protein (Wells and Cordingley, 1992), although the latter notably lacks proline (Fig. 5). Curiously, all these proteins have one thing in common: they are significant components of structures that function in tension.

Acknowledgments

We thank Alan Jordan for year-round collections of *D. polymorpha* and the National Sea Grant Program of NOAA for support. Drs. John McDonald and Alison Hunt provided generous assistance with DNA sequencing. Dr. Lesz Rzepecki generously provided samples of native Dpfp1, and Luis Burzio provided assistance and advice regarding analysis of Dpfp1 by mass spectrometry. Research was supported by grants from the National Oceanic and Atmospheric Administration (94-3501-0115) and the Office of Naval Research (N00014-96-1-1205) to JHW. KEA was supported in part by USPHS Grant T32-GM08550. cDNA sequences have been submitted to GenBank (Accession # AF043221; AF043222; AF043223).

Literature Cited

Ackerman, J. D., C. R. Ethier, D. G. Allen, and J. K. Spelt. 1992. Investigation of zebra mussel adhesion strength using rotating disks. *J. Environ. Eng.* **118:** 708–724.

Akashi, H. 1994. Synonymous codon usage in *Drosophila melanogaster:* natural selection and translational accuracy. *Genetics* **135:** 927–935.

Chomczynski, P., and N. Sacchi. 1987. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal. Biochem.* **162:** 156–159.

Coyne, K. J., X. X. Qin, and J. H. Waite. 1997. Extensible collagen in mussel byssus—a natural block-copolymer. *Science* **277:** 1830–1832.

Eddington, N. D. 1996. *Partial oligonucleotide sequence of a mussel byssal precursor protein,* Dreissena polymorpha *foot protein 2.* M. S. thesis, University of Delaware, Lewes.

Ertl, H., A. Hallmann, S. Wenz, and M. Sumper. 1992. A novel extensin that may organize extracellular matrix biogenesis in *Volvox carteri*. *EMBO J.* **11:** 2055–2062.

Filpula, D. R., S. M. Lee, R. P. Link, S. L. Strausberg, and R. L. Strausberg. 1990. Structural and functional repetition in a marine mussel adhesive protein. *Biotechnol. Prog.* **6:** 171–177.

Frohman, M. A., M. K. Dush, and G. R. Martin. 1988. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proc. Natl. Acad. Sci. USA* **85:** 8998–9002.

Hong, J. C., R. T. Nagao, and J. L. Key. 1987. Characterization and sequence analysis of a developmentally regulated putative cell wall protein gene isolated from soybean. *J. Biol. Chem.* **262:** 8367–8376.

Inoue, K., and S. Odo. 1994. The adhesive protein cDNA of *Mytilus galloprovincialis* encodes decapeptide repeats but no hexapeptide motif. *Biol. Bull.* **186:** 349–355.

Inoue, K., Y. Takeuchi, D. Miki, and S. Odo. 1995. Mussel adhesive plaque protein gene is a novel member of epidermal growth factor-like gene family. *J. Biol. Chem.* **270:** 6698–6701.

Inoue, K., Y. Takeuchi, D. Miki, S. Odo, S. Harayama, and J. H. Waite. 1996a. Cloning, sequencing and sites of expression of genes for the hydroxyarginine-containing adhesive-plaque protein of the mussel *Mytilus galloprovincialis*. *Eur. J. Biochem.* **239:** 172–176.

Inoue, K., Y. Takeuchi, S. Takeyama, E. Yamaha, F. Yamazaki, S. Odo, and S. Harayama. 1996b. Adhesive protein cDNA sequence of the mussel *Mytilus coruscus* and its evolutionary implications. *J. Mol. Evol.* **43:** 348–356.

Johnson, L. E., and D. K. Padilla. 1996. Geographic spread of exotic species: ecological lessons and opportunities from the invasion of the zebra mussel *Dreissena polymorpha*. *Biol. Conserv.* **78:** 23–33.

Kafatos, F. C., N. Spoerel, S. A. Mitsialis, H. T. Nguyen, C. Ramano, J. R. Lingappa, B. D. Mariani, G. C. Rodakis, R. Leganidou, and S. G. Tsitilou. 1987. Developmental control and evolution in the chorion gene families of insects. *Adv. Genet.* **24:** 223–242.

Kieliszewski, M. J., and D. T. A. Lamport. 1994. Extensin: repetitive motifs, functional sites, post-translational codes, and phylogeny. *Plant J.* **5:** 157–172.

Kozak, M. 1986. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* **44:** 283–292.

Labeit, S., and B. Kolmerer. 1995. Titins: giant proteins in charge of muscle ultrastructure and elasticity. *Science* **270:** 293–296.

Laursen, R. A. 1992. Reflections on the structure of mussel adhesive proteins. Pp. 55–74 in *Structure, Cellular Synthesis and Assembly of Biopolymers*, S. T. Case, ed. Springer Verlag, Berlin.

Mita, K., S. Ichimura, and T. C. James. 1994. Highly repetitive structure and its organization of the silk fibroin gene. *J. Mol. Evol.* **38:** 583–592.

Nielsen, H., J. Engelbrecht, S. Brunak, and G. von Heinjne. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10:** 1–6.

Papov, V. V., T. V. Diamond, K. Biemann, and J. H. Waite. 1995. Hydroxyarginine-containing polyphenolic proteins in the adhesive plaques of the marine mussel *Mytilus edulis*. *J. Biol. Chem.* **270:** 20183–20192.

Patwary, M. U., M. E. Reith, and E. L. Kenchington. 1996. Isolation and characterization of a cDNA encoding an actin gene from sea scallop (*Placopecten magellanicus*). *J. Shellfish Res.* **15:** 265–271.

Qin, X. X., and J. H. Waite. 1995. Exotic collagen gradients in the byssus of the mussel *Mytilus edulis*. *J. Exp. Biol.* **198:** 633–644.

Qin, X. X., K. J. Coyne, and J. H. Waite. 1997. Tough tendons: mussel byssus has collagen with silk-like domains. *J. Biol. Chem.* **272:** 32623–32627.

Roberts, L. 1990. Zebra mussel invasion threatens U.S. waters. *Science* **249:** 1370–1372.

Robinson, M., R. Lilley, S. Little, J. S. Emtage, G. Yamamoto, P. Stephens, A. Millican, M. Eaton, and G. Humphreys. 1984. Codon usage can affect efficiency of translation of genes in *Escherichia coli*. *Nucleic Acids Res.* **12:** 6663–6671.

Rzepecki, L. M., and J. H. Waite. 1993. The byssus of the zebra mussel, *Dreissena polymorpha*. II: structure and polymorphism of byssal polyphenolic protein families. *Mol. Mar. Biol. Biotechnol.* **2:** 267–279.

Sambrook, S., E. F. Fritsch, and T. Maniatis. 1989. *Molecular Cloning: A Laboratory Manual.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Sorensen, M. A., C. G. Kurland, and S. Pedersen. 1989. Codon usage determines translation rate in *Escherichia coli*. *J. Mol. Biol.* **207:** 365–377.

Taylor, S. W., D. B. Chase, M. H. Emptage, M. J. Nelson, and J. H. Waite. 1996. Ferric ion complexes of DOPA-containing adhesive protein from *Mytilus edulis*. *Inorg. Chem.* **35:** 7572–7577.

Varenne, S. J., J. Buc, R. Lloubes, and C. Lazdunski. 1984. Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. *J. Biol. Chem.* **180:** 549–576.

von Heijne, G. 1985. Signal sequences: the limits of variation. *J. Mol. Biol.* **184:** 99–105.

Waite, J. H. 1990. The phylogeny and chemical diversity of quinone-tanned glues and varnishes. *Comp. Biochem. Physiol.* **97B:** 19–29.

Waite, J. H., T. J. Housley, and M. L. Tanzer. 1985. Peptide repeats in a mussel glue protein: theme and variation. *Biochemistry* **24:** 5010–5014.

Wells, K. E., and J. S. Cordingley. 1992. The cell and molecular biology of eggshell formation in *Schistosoma mansoni*. Pp. 97–114 in *Structure, Cellular Synthesis and Assembly of Biopolymers*, S. T. Case, ed. Springer Verlag, Berlin.